DOCUMENT RESUME

EA 002 486

ED 031 814

Proceedings of AERA Symposium: The World of Evaluation Needs Reshaping (Los Angeles, California, February 8, 1969).

Northwest Regional Educational Lab., Portland, Oreg.
Pub Date Feb 69
Note-71p.; Papers presented at the AERA Symposium.
EDRS Price MF-$0.50 HC-$3.65
Descriptors-*Decision Making, Educational Change, *Educational Innovation, Evaluation Criteria, *Evaluation Methods, Evaluation Techniques, Information Theory, *Program Evaluation, *Theories
Identifiers-ESEA Title 1, ESEA Title 3

The Symposium proceedings point out that the evaluation of educational innovations awaits the modernization of evaluation theory. Specific approaches to the problem are presented in five papers, as follows: (1) "An Overview of the Evaluation Problem," by Egon G. Guba, Associate Dean, School of Education, Indiana University, Bloomington, Indiana; (2) "An Emergent Theory of Evaluation," by Daniel L. Stufflebeam, Director, Evaluation Center, Ohio State University, Columbus, Ohio; (3) "Knowledge About Decision Processes and Information," by Robert S. Randall, Director, Division of Program Research and Evaluation, Southwest Educational Development Laboratory, Austin, Texas; (4) "Evaluation Designs and Instruments," by Jack C. Merwin, Director of Psychological Foundations, College of Education, University of Minnesota, Minneapolis, Minnesota; and (5) "The World of Evaluation Needs Reshaping," by Michael C. Giammatteo, Research and Development Specialist, Northwest Regional Educational Laboratory, Portland, Oregon. (JK)

ED031814

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

PROCEEDINGS OF

AERA SYMPOSIUM: THE WORLD OF EVALUATION NEEDS RESHAPING

Hilton Hotel
Los Angeles, California
February 8, 1969

EA 002 486

Reproduced and Made Available
With Permission From the Authors

By

Northwest
Regional
Educational
Laboratory

710 S.W. Second Avenue
400 Lindsay Building
Portland, Oregon 97204
Telephone (503) 224-3650

AERA SYMPOSIUM

A report of the proceedings of the AERA Symposium entitled "The World of Evaluation Needs Reshaping." Acknowledgement is made to the participants who have all graciously agreed to allow their papers to be reproduced and thus made available to a larger audience.

ABSTRACT

Traditional methods of evaluation have failed educators as they have sought to assess the impact of innovations in operating programs. For years the evidence produced by the application of conventional evaluation procedures has contradicted the experiential evidence of the practitioner. Educational innovations have persisted despite the lack of supporting evidence.

The evaluation efforts mounted in relation to new federally-funded programs continue to be relatively fruitless and quite independent of the utility of existing evaluation techniques. Why cannot the educational community respond to the urgent need for useful evaluative information? Why cannot evaluation programs be designed and implemented that will quickly solve these problems?

The participants in this symposium believe that the situation cannot be explained simply on the grounds of ignorance, carelessness or unconcern. They believe it exists because there are certain crucial lacks:

1. Lack of adequate evaluation theory.
2. Lack of knowledge about decision processes and information requirements.
3. Lack of instruments and designs.
4. Lack of mechanisms for organizing and reporting evaluative information.
5. Lack of trained personnel.

These five lacks pose a formidable challenge to the educational community. Even the best evaluators can function only with extant theory, concepts, designs, tools, mechanisms, and training. The educational

practitioner can hardly be blamed if, when placing his faith on those extant procedures and consultant help, he produces evaluative results of little use to anyone. Nor can we fault him if he becomes disenchanted with the substitutes we offer because they are not operational.

The primary task in evaluation today is the provision of sensible alternatives to the evaluator. The evaluation of educational innovations awaits the modernization of evaluation in education.

The purpose of this symposium was to focus the attention of educational researchers on the evaluation dilemma and to generate a wider community of interest in attacking this situation.

Participants

Symposium Chairman
Ray E. Jongeward
Director of Research and Evaluation
Northwest Regional Educational Laboratory
Portland, Oregon

An Overview of the Evaluation Problem
Egon G. Guba
Associate Dean
School of Education
Indiana University
Bloomington, Indiana

An Emergent Theory of Evaluation
Daniel L. Stufflebeam
Director, Evaluation Center
The Ohio State University
Columbus, Ohio

Knowledge About Decision Processes and Information
Robert S. Randall
Director, Division of Program Research and Evaluation
Southwest Educational Development Laboratory
Austin, Texas

Abstract (Continued)

Evaluation Designs and Instruments
Jack C. Merwin
Director of Psychological Foundations
College of Education
University of Minnesota
Minneapolis, Minnesota

The World of Evaluation Needs Reshaping
Michael C. Giammatteo
Research and Development Specialist
Northwest Regional Educational Laboratory
Portland, Oregon

AN OVERVIEW OF THE EVALUATION PROBLEM

EGON G. GUBA
Associate Dean
School of Education
Indiana University

A Paper read at the Symposium on "The World of Evaluation Needs Reshaping," Los Angeles, American Educational Research Association Convention, February, 1969

Introduction

The American educational establishment is currently making a massive effort at self-improvement. Unprecedented resources, stemming mainly from the Federal Government under the provisions of ESEA but coming also from foundations, state departments of education, local school systems, private industries, and other agencies, are being expended on a variety of promising but as yet unproved programs. To assure the effective and efficient uses of these resources, and even more importantly, to determine the real utility of the innovative approaches, it is necessary to gather hard data about their performance. Evaluation is the process best suited for this purpose.

The traditional methods of evaluation have, however, failed educators as they have sought to assess the impact of innovations in operating systems. Indeed, for decades the evidence produced by the application of conventional evaluation procedures has contradicted the experiential evidence of the practitioner. Innovations have persisted in education not because of the supporting evidence of evaluation but despite it. A recent dramatic example is afforded by the Higher Horizons program in New York City. Test data failed to affirm what supervisors, teachers, and clients insisted was true--that the program was making a difference so great that it simply could not be abandoned.

On a broader scale, the recent Coleman report circulated by the Office of Education has shocked educators by noting that "one implication stands out above all: that schools bring little influence to bear on a child's

achievement that is independent of his background and general social context."[1] More specifically Coleman states that there is a " . . . relatively small amount of school-to-school variation that is not accounted for by differences in family background, indicating the small independent effect of variations in school facilities, curriculum, and staff on achievement."[2]

This conclusion is incredible on its face. It means, if true, that it makes little difference whether a teacher is good or bad, whether good or poor materials are available, or whether the school is a barn or a geodesic dome; students will learn about the same. Now anyone who has spent any time at all in a school knows that is just not so; why then do our evaluative techniques fail to detect the effects?

When the evidence produced by any scientific concept or technique continually fails to affirm experiential observation and theory arising from that observation, the technique may itself appropriately be called into question. It shall be the burden of my remarks that evaluation as we know it has failed, and that the world of evaluation does indeed require, as the title of this symposium suggests, reshaping.

Some Clinical Signs of Failure

Can this contention of failure really be supported? Let us look at some of the clinical signs that present evaluation is somewhat less than effective:

---

[1]James S. Coleman et al., Equality of Educational Opportunity. National Center for Educational Statistics, U. S. Government Printing Office, Washington D.C., 1966, p. 325.

[2]Ibid.

1. Avoidance. A certain sign of evaluation's failure is that everyone avoids it unless it becomes painfully necessary. This tendency toward avoidance can be noted at all levels. Local school districts rarely incorporate evaluation into any effort which they themselves fully control and finance. This is particularly clear when one consults proposed project budgets, if evaluation costs are included at all they are contemplated only in very general terms i.e., perhaps the salary of an evaluation "expert," or the cost of buying commercially available instruments.

The same avoidance is evident within state departments and even within the U. S. Office of Education, which, despite a great deal of talk about the desirability of evaluation for the schools, never budgets or staffs sufficiently well to provide for evaluation of its own programs.

2. Anxiety. The psychiatrist is very familiar indeed with the phenomenon of "free-floating" anxiety, which characterizes many neurotic patients. A similar affliction characcerizes the practitioner and the professional evaluator when they approach an evaluation. This anxiety seems to stem from the ambiguities of the evaluation process. Since so many elements of that process are badly understood, the particular evaluation that may be applied may yield random, meaningless data. And who is there among us that would not feel anxious if judgments were to be made about our programs, our decisions, or our effectiveness by what may be a random process? Our protests that no truly professional practitioner need feel anxious when confronted by the need to evaluate are empty and worthy of contempt.

3. Immobilization. Despite the opportunity that has existed for four or more decades, schools have not responded to evaluation in any meaningful way; indeed, the mere existence of an office or functionary within the schools charged with systematic evaluation is still rare. Further, despite the federal requirements for evaluation built into legislation, particularly Titles I and III of ESEA, evaluative data are still relatively non-existent, as are programs that could be pointed to as "models" of what might be done in evaluation. This lethargy, this lack of responsiveness, this immobilization can only be taken as symptomatic of a deeper illness.

4. Lack of guidelines. The lack of meaningful and operational guidelines for evaluation is notable. Consider for example the statement made in the ESEA Title III manual published by the U. S. Office of Education:

> A. Where applicable, describe the methods, techniques, and procedures which will be used to determine the degree to which the objectives of the proposed program are achieved.
>
> B. Describe the instruments to be used to conduct the evaluation.
>
> C. Provide a separate estimate of costs for evaluation purposes. This amount should be included in the proposal budget summary.[3]

While these three statements are expanded with some 2½ pages of text, the expansion does little to inform the reader about anything other than technical requirements. The guidelines are subject to very wide interpretation and offer little operational assistance to the proposal developer.

---

[3] A Manual for Project Applicants and Grantees (Title III Elementary and Secondary Education Act). Washington, D. C., Office of Education, Department of Health, Education, and Welfare, May, 1967, p. 48.

The inability of the very agencies that require evaluation to provide adequate guidelines for its implementation must be regarded as one of the more serious difficulties besetting evaluation.

5. Misadvice. Evaluation consultants, many of whom are drawn from the ranks of methodological specialists in educational research, fail to give the kind of advice which the practitioner finds useful. Indeed, the practitioner may be led down the primrose path instead. A recent analysis of a small sampling of Title III proposals gives the flavor of this difficulty.[4] Twenty-one proposals were examined, but only one was found that could be considered to have an adequate design from a traditional methodological point of view. Most had no design at all, while those that did offered designs well known to suffer from serious deficiencies. Yet the majority of these 21 proposals purported that the services of an evaluation specialist had been employed and that he was primarily responsible both for the planning and the implementation of the evaluation program. Usually the consultant and his institutional affiliation were named so that there was no doubt about his technical competence. It is certainly a serious symptom of disorder when the experts in the field of evaluation seem to be unable to design evaluations that meet even their own criteria of technical soundness.

6. No significant differences. Another very significant indication that evaluation is in trouble is the fact that it is so often incapable of uncovering any significant information. Over and over comparative studies

---

[4]Egon G. Guba, "Report on the Evaluation Provisions of Twenty-One Title III Proposals," Report to the National Panel on Title III Evaluation, Richard I. Miller, Director, October 15, 1967.

of alternatives in education have ended in a finding of "no significant difference." Several conventional responses are made to this situation. It is often observed that the educationists are incapable of devising any approaches that are better than those things that they are already doing. But if this is so we ought perhaps to applaud their remarkable consistency, since they do not devise alternatives that are any worse either! Another oft heard response is to say that the lack of efficacy of comparative studies is well established by this consistent failure to find differences; educationists are then warned not to engage in such studies because to do so is to behave stupidly. This equally glib response of course ignores the fact that this comparative question is exactly the one that must be asked if improvement is to occur. What could be more relevant, as one gropes to change for the better, than to ask about alternatives and to seek to determine which of several available alternatives, including present practice, is most efficacious?

This brief listing of the most obvious clinical signs of evaluation's failure is compelling. Any professional area that is so much avoided; that produces so many anxieties; that immobilizes the very people who want to avail themselves of it; that is incapable of operational definition even by its most trained advocates, who in fact render bad advice to the practitioners who consult them; which is not effective in answering reasonable and important questions and which has made little apparent effort to isolate and ameliorate its most serious problems; must indeed give us pause.

The Basic Lacks

How can one account for this state of affairs? Why cannot the educational community respond to the urgent need for useful evaluative information? Why cannot evaluation programs be designed and implemented that will quickly eradicate this shortage of data? The situation cannot be explained simply on the grounds of ignorance, carelessness, or unconcern. It exists because of certain crucial lacks:

1. Lack of adequate definition of evaluation. Evaluation, like any analytic term, can be defined in many essentially arbitrary ways. Each of the ways which have gained common acceptance have certain utilities and certain disadvantages.

An early definition of evaluation tended to equate that term with measurement, as it had developed in the twenties and thirties. We must remember that historically, the evaluation movement followed upon the heels of, and was made technically feasible by, the measurement movement. The technique of equating a new movement with an older established movement in order to gain credibility is common, as for example, in calling "social science" a science in order to gain some of the status reserved in this society for a scientific venture. Moreover, the instrumentation developed by measurement experts provided the conceptual basis for evaluation. Finally, and perhaps most important, the use of measurement devices resulted in scores and other indices that were capable of mathematical and statistical manipulation, which in turn rendered possible the handling of masses of data and the easy comparison of individual or classroom scores with group norms. Thus

the idea of interpreting evaluative data in relation to an objective criterion could be introduced, but the criterion (norms) was devoid of value judgments and was, sociologically and culturally, antiseptic.

What disadvantages accrue from such a definition? First, evaluation was given an instrumental focus; the science of evaluation was viewed as the science of instrument development and interpretation. Second, the approach tended to obscure the fundamental fact that value judgments are necessarily involved (a problem to which we shall return below). Third, evaluation tended to be limited to those variables for which the science of measurement had successfully evolved instruments; other variables came to be known as "intangibles," a characterization which was equivalent to saying that they couldn't be measured; hence had no utility, and ultimately, no importance. Thus the limits placed upon evaluation because of a lack of instrumental sophistication came to be viewed as the real limits to which evaluation had to be constrained. In short, this definition results in an evaluation which is too narrow in focus and too mechanistic in its approach.

Another definition of evaluation which has had great currency is that of determining the congruence between performance and objectives, especially behavioral objectives. This congruence definition, which grew out of the work of Tyler and others at Ohio State University, particularly in connection with the Eight Year Study, had an enormous impact on education, as well it might. In the first place, the definition appeared in connection with an organized rationale about the entire instructional process, and provided a means whereby the teacher, administrator, supervisor, and curriculum maker could make sensible judgments about what they were doing. Evaluation no longer focussed solely on

the student, but could provide insights about the curriculum and other educational procedures as well. The utility of evaluation was thus broadened and for the first time, a practical means was devised to provide feedback (a term unheard of at the time). Finally, evaluation came to have utility not only for judging a product (student achievement, for example) but also a process (the means of instruction, for example), a distinction whose import is only now being fully realized.

What disadvantages accrue as a result of this definition? First, with the heavy emphasis that this approach placed on objectives, the major task of the evaluator came to be seen as developing a set of objectives that were sufficiently operational so that the required congruence assessment could occur. The objectives themselves, in general form, were obtained by an almost mystic process that remained relatively unspecified; Tyler spoke eloquently about "screening objectives through a philosophy and a psychology," but these were vague terms. The real problem was to take the general "screened" objectives and by a process of successively finer definition and expansion reduce them to their most operational form.

A second disadvantage of this approach was the fact that the objectives were to be stated in behavioral terms. A "true" evaluation could take place only by reduction to student behaviors. Thus we are confronted with such absurdities as trying to evaluate the effectiveness of a new staff recruitment procedure, for example, by showing that this somehow related to increased achievement on the part of students.

A third and perhaps major disadvantage of this approach is that the emphasis on student behavior as the criterion caused evaluation to become

a post facto or terminal technique. Data became available only at the end of a long instructional period. It is perhaps ironic that a definition that hinted so clearly at feedback and its utilization in improvement should have this effect. The full possibilities were thus not only not realized but the form of the definition froze evaluation as a terminal event rendering product judgments. If process data were available they could only be utilized the next time round; it was too late to use them for refinement in the ongoing program, i.e., in the program from which the evaluative data were extracted.

Thus, the definition of evaluation in congruence terms relating outcomes to objectives, while broadening the utility of evaluation considerably and providing the possibility for feedback and process data, did tend to label evaluation as a terminal process that yielded information only after the fact.

Neither of the two previously discussed definitions of evaluation placed much emphasis on the judgmental process. Certainly in the case of the measurement definition, and to some extent in the case of the congruence definition, the matter of placing value on the data was, if considered at all, taken pretty much for granted. But there was a school of thought, entertained mainly by persons who would not have labeled themselves as evaluators, that defined evaluation in yet a third way, viz., that evaluation is professional judgment. Perhaps the most obvious example of this definition is in the visitation procedure used by the various accrediting associations such as the North Central Association. While evaluative criteria do exist, these are applied mainly by school personnel whose school is being evaluated, not by the visitation teams. The chief value in their application is often

understood to be the process of application rather than the results obtained thereby; the school personnel through this exercise gain new insights into themselves, their problems, and their shortcomings. The actual evaluations are made not by the school personnel, however, but by the visitation teams, who come in, "soak up" the data by virtue of their expertise and experience, and render a judgment. The judgment is the evaluation.

A similar approach can be seen in the traditional school survey, and in the use of panels by the Office of Education, Foundations, and other funding agencies to evaluate proposals. Again, the evaluation is whatever judgment they render.

Advantages of this approach are fairly obvious. First, the evaluation is quickly managed. Second, the evaluators are typically experts with a great deal of experience which they can bring into play without being artifically constrained by "instruments." Third, the interplay of a variety of factors in a situation is taken into account more or less automatically, and the evaluator is thus freed of the problem of relating and aggregating data after he has collected them. Finally, there is no appreciable lag between data collection and judgment; we do not need to wait for long time periods while data are being processed.

Despite these apparent advantages, however, there are very few people who would willingly rely on this approach unless nothing else can be done. First, one has the feeling that it is not so much a matter of convenience but of ignorance that forces such an approach; if we knew more we could be more precise and objective. Secondly, we have fears for the reliability and the objectivity of such judgments, and how can one demonstrate

whether they are or are not reliable and objective? It is this inability to apply the ordinary prudent tests of scientific inquiry that makes us leery, even when we are willing to concede the expertness of the evaluators involved. Third, the process hides both the data considered and the criteria or standards used to assess them, because the process is implicit. Thus, even if the judgments are valid, reliable, and objective, we have little confidence that we can tell why they are so, or to generalize to other situations. Thus, to sum up, the inherent uncertainty and ambiguity of evaluations based on this definition leave one dissatisfied.

It is apparent from this review of common definitions of evaluation that while each definition offers the evaluator certain advantages, each is also accompanied by certain disadvantages. No definition is available that does not have several serious disadvantages as concomitants.

2. Lack of adequate evaluation theory. There have been, for all practical purposes, no advances in the theory of evaluation since Ralph Tyler completed his formulations during the decade of the forties. Since that time the professionals in the field have felt content simply to borrow from the methodology of other fields, notably educational research. Indeed, the methodology of education has come to be equated with the methodology of research, with disastrous consequences. Let us examine some of these:

a. Laboratory antisepsis. The purpose of research is to provide new knowledge. Its methodology is designed to produce knowledge which is universally valid. The purpose of a laboratory is to provide such a context-free environment, within which universally true knowledge can be developed. The establishment of close controls makes it possible to rule out all influences except those which are the object of inquiry.

Evaluations are not designed to establish universal laws, however, but to make possible judgments about some phenomenon. In this situation one not only does not want to establish highly controlled conditions in which possible sources of confounding are filtered out, but in fact one wishes to set up conditions of invited interference from all factors that might ever influence a learning (or whatever) transaction.

Thus, educational evaluation does not need the antiseptic world of the laboratory but the septic world of the classroom and the school in order to provide useful data. The use of laboratory research designs and techniques poses conditions that are simply inappropriate for the purposes for which one does an evaluation.

b. The effects of intervention. The interest of a researcher, particularly in the laboratory, is usually focussed on the interplay of certain so-called independent and dependent variables. The researcher must engage in some form of manipulation or intervention to arrange for the conditions necessary to study this interaction. Thus the investigator becomes an integral part of the data since they would not have occurred without his presence.

By intervening in a situation an investigator can achieve the controls necessary to allow him to focus upon segments and processes of particular concern to him. But he does this at a possible loss of information because he is dealing with a contrived situation. It is also possible, however, to collect data which are natural and uncontrived, but which are also uncontrolled, difficult to analyze, and of course which allow all factors to exert whatever influence they might. It is about such actual

situations that the evaluator wants information, not the contrived situations which, regardless of their utility for other purposes (e.g., establishing universally true principles) are not appropriate for the evaluator's purpose.

c. Terminal availability. The typical research design is concerned with the assessment of the effects of some "treatment" or combination of treatments. A major intent of design is to arrange matters so that the influence of factors not included in the treatment(s) are either controlled or randomized while the effect of the treatment is being detected. At the end of some period of time sufficient for the treatment to produce its presumed effect measures are taken from which a judgment can be drawn.

This general format produces data only at the termination of the experiment. If the treatment is judged, let us say, to have been inappropriate or insufficient, nothing can be done to improve the situation for the test subjects from whom the insufficiency was judged. But suppose that the intent had been, as it often is in the case of education, to improve the treatment while it was being applied, so that the maximum benefit might be derived not only for the future but also for the group on which the experiment was conducted. When we try a new method of reading for disadvantaged children we are just as interested in the children we try it on as we are in other children who may use it in the future. The evaluator cannot be content with terminal availability. The traditional methodology will not help him.

d. Single evaluations only. Evaluators operating on the basis of classic research methodology must insist, for the sake of control, that no more than one evaluation be conducted simultaneously, lest one confound the other. It is impossible, using such an approach, to distinguish the effects

of two new treatments being evaluated simultaneously, at least not without very expensive refinements. But again, moral principles prevent the educator from keeping the possible benefits of a new treatment from a group of children just because they are already being exposed to another treatment designed to remedy some other problem.

e. The inapplicability of assumptions. Classical research methodology and the statistical analyses which are appropriate thereto are based upon a series of assumptions which do not meet evaluation requirements too well.

There are first of all the assumptions underlying the statistical techniques. Normality of distribution, for example, is necessary to make even certain descriptive statistics meaningful, such as that the interval included between the mean plus and minus one standard deviation shall include 68 per cent of the cases. Other assumptions are built into the interpretive tables in which the "significance" of analytic statistics is determined; thus the derivation of the F distribution depends upon certain random sampling assumptions. Finally still other assumptions are necessary to support the logical derivation of the interpretive techniques; thus, in the case of analysis of variance, the additivity assumption which asserts that treatments have equal effects on all persons to whom they are applied is vital. None of these assumptions is likely to hold in typical evaluation situations. To cite one example, it is clear that good teaching tends to interact with pupils so that the able learn more than the less able. The additivity assumption thus is very tenuous.

It is well known that statistical techniques are "robust" with respect to those assumptions, that is, the statistics tend to provide valid information even though the assumptions may be rather sharply violated. Nevertheless it is one thing simply to deviate from certain assumptions and quite another to attempt to apply techniques in situations where their assumptions are patently not met. Even the most robust of techniques might be adversely affected if enough of its assumptions were systematically violated.

f. The impossibility of continuous refinement. Perhaps the most damaging assertion that may be made about the application of conventional experimental design to evaluation situations is that such application conflicts with the principle that evaluation should facilitate the continuous improvement of a program. Experimental design prevents rather than promotes changes in the treatments because, as has been noted, treatments cannot be altered if the data about differences between treatments are to be unequivocal. Thus, the treatment must accommodate the evaluation design rather than vice versa. It is probably unrealistic to expect directors of innovative projects to accept these conditions. Obviously, they cannot constrain a treatment to its original, undoubtedly imperfect form just to ensure internally valid end-of-year data. Rather, project directors must use whatever evidence they can obtain continuously to refine and sometimes radically to change both the design and its implementation. Concepts of evaluation are needed which would result in evaluations which would stimulate rather than stifle dynamic development of programs. Clearly, equating evaluation methodology with research methodology is absolutely destructive of this aim.

3. Lack of knowledge about decision processes. Programs to improve education depend heavily upon a variety of decisions, and a variety of information is needed to make and support those decisions. Since the purpose of evaluation is to provide this information, the evaluator must have adequate knowledge about the relevant decision processes and associated information requirements before he can design an adequate evaluation. At present no adequate knowledge of decision processes and associated information requirements relative to educational programs exists. Nor is there any ongoing program to provide this knowledge.

A first question that must be considered is what model of the decision-making process is most productive for evaluators to have in mind. Most treatises on the subject of decision-making view the process as essentially rational: the decision-maker starts with some awareness of a problem which he must resolve; he then assembles alternative ways of responding to that problem; he chooses from among the alternative responses that one which, on balance, appears to have the highest success probability, and then he implements the choice.

But it seems highly unlikely that real-world decisions are in fact made in these ways. The mere creation of awareness of the need for a decision is a formidable task; many decision-makers seem to prefer not to be made aware unless absolutely necessary. Generally speaking, the range of possible responses available to the decision-maker is not very large; if even one alternative exists the decision-maker is usually delighted. The choice among alternatives, is not usually made on the basis of explicit and well-understood criteria; many decision-makers pride themselves on "shooting from the hip" and would not have it any other way.

Attempts have been made to define other models of the decision-making process; a notable example is the model of "disjointed incrementalism" proposed by Braybrooke and Lindblom.[5] It is likely that such other models may have more utility for the evaluator than the conventional rational model. But meantime, it is clear that evaluators have not had a clear and useful conception in mind, a fact which has hindered them considerably in determining what evaluation methodologies are most productive and what kinds of information delivered under what circumstances would be most valuable.

A second problem relating to decision-making is the lack, to date, of adequate taxonomies of educational decisions. If evaluation is to serve decisions it would be most useful indeed to be able to categorize or classify educational decisions by type so that, for example, evaluation designs appropriate to each type might be conceptualized. But what is the range and scope of educational decision-making? What substantive concerns are reflected in these decisions?

A third problem is the lack of methodologies for linking evaluation to the decision-maker whom it is ultimately to serve. One such linkage problem has already been alluded to--that of creating awareness in the decision-maker of the need for a decision. Another is that of helping the decision-maker to identify the criteria which he is using or might use--a difficult matter which implies a professional relationship of the highest order between evaluator and client. A third aspect has to do with reporting evaluative

---

[5]David Braybrooke and Charles E. Lindblom, A Strategy of Decision: New York: The Free Press, 1963.

information to the decision-maker in ways which he finds credible and helpful. The evaluator is often thought of as a high level technician familiar with the methodologies of research and data analysis, but it is clear that in dealing with the decision-maker he plays a series of professional roles more similar to those of the counselor or attorney than to the educational researcher. Methodologies for this role are simply lacking.

4. Lack of criteria. Most evaluators agree that the mere collection of data does not constitute evaluation--there is always at least a hint of making judgments about the data in terms of some implicit or explicit value structure. Thus it would be unusual to speak just about whether or not objectives are achieved, but rather how well they are achieved. The need to introduce values gives rise to a number of problems. First there is the matter of where the values come from. It was pointed out that scholars who defined evaluation as the congruence between performance and objectives paid little attention to the origin of the objectives except that they were to be "screened" through a psychology and a philosophy. This doctrine leaves untouched the question of what philosophy and what psychology should be used as screens. When this question is made explicit it is quickly apparent that no adequate methodology exists for the determination of values, even though, as we have already implied, such a determination may constitute the most professional task which the evaluator performs. It may, indeed, be his chief claim to a professional rather than a technical role.

Another question that arises in this domain is how to achieve consensus about the values that are to be invoked in evaluations. It may be fairly easy to achieve consensus at a micro level, as for example, when a

group of English teachers attempts to define what the objectives shall be for the freshman composition course. But how can one achieve consensus on the purposes of ESEA Title I? How is one to interpret evaluative data to meet the value standards that might be invoked? In a pluralistic society in which multiple values necessarily exist side-by-side, which values will be served? Indeed, how can one even determine what the value patterns are? And finally, when such multiple values are applied, will it not almost inevitably be the case that the same data when interpreted in terms of different value standards will give rise to antithetical evaluations?

Finally, there is a variant of the value problem which concerns the values of the evaluators themselves, and which accounts for at least some of the apparent estrangement between the evaluator and the practitioner. The practitioner must necessarily take a variety of considerations into account when he makes any decision. At times he may find economic considerations most compelling, or political ones. But the evaluator is much more inclined to adhere, almost exclusively at times, to so-called scientific values. He prefers to make his decisions on "hard" data, by which of course he means scientifically derived data. Since he prides himself on being "rational," he cannot understand why everyone else is not rational too. He feels disinclined to apply his scientific methods to a determination, say, of what the political climate is, because to do so would prostitute himself and pervert the ideals of the scientific community. This estrangement is severe and cannot be lightly dismissed.

5. Lack of approaches differentiated by levels. The problem of levels, as the term will be used here, stems from the fact that the evaluator's traditional point of focus has been microscopic, e.g., the individual student, the classroom, or the school building, rather than macroscopic, e.g., the school district, the state system, or the national network. This microscopic focus serves the evaluator badly when he is confronted with evaluation problems at superordinate levels, as is often the case today.

One consequence of this misplaced focus is that the techniques the evaluator uses are inappropriate. An example we have already noted is that at the macroscopic level, it makes little sense to focus on behavioral objectives. Another difficulty is that the instruments have been developed for use with individuals, while the evaluator may now be concerned with system data. Finally, the evaluator is usually concerned with all of the subjects at the micro level, e.g., all of the students taking a certain science course in a certain school, while at the macro level he must lean heavily on sampling procedures with which he is not too familiar or which remain to be developed to an acceptable technical degree, as for example, using item sampling procedures rather than having all of the students answer all of the test items.

Another consequence is faulty aggregation, which takes two forms. First, there is the matter of summarizing operational data obtained at micro levels. Clearly the amount and kind of information required by the local project simply jams the wheels at the macro level. The second form of the aggregation problem is, in a sense, the inverse of the first; while these reports of operational data may more than meet the requirements of the micro agency

they do not contain information which is of vital concern to the macro agency. Thus the local agency will not collect data relevant to the question of, say, how the Title III program is doing as a whole, while overloading the macro agency with information about how the specific project is doing. Overall, this aggregation problem seems often to be a matter of too much of the wrong thing.

A third consequence is that of conflicting purposes. Different data or information may be required at different levels, as well as different criteria to assess them. The purposes of agencies at different levels vary markedly. While there may be little question that the purpose of the teacher is to teach, or that the success of her teaching may be most appropriately assessed by reference to some criterion relating to student achievement, it is equally true that this purpose and this criterion are not relevant to, say, the evaluation of a statewide supervision or program or a national curriculum improvement effort.

Thus, the introduction of various levels of evaluation introduces problems that are by no means able to be resolved through the application of techniques, methods, criteria, and perspectives developed at the micro level, where we are accustomed to working. This fact must be recognized and steps must be taken to develop the new approaches that are clearly required. Evaluators must learn to "think big," and thinking big involves more than a quantitative increase in perspective.

6. Lack of mechanisms for organizing, processing, and reporting evaluative information. Even if the above lacks did not exist, there still would remain an important logistical problem related to organizing, processing,

and reporting evaluative information. There is no central, coordinated, comprehensive system of educational data processing, storage, and retrieval in existence. A few prototypes may be noted, one at the University of Iowa, but these prototypes do not begin to encompass the masses of information which will need to be processed. Meantime, one must count on the archaic and usually different systems employed by the various school systems and state departments of education.

7. Lack of trained personnel. Evaluation personnel have always been in short supply in this country, but the new improvement programs have magnified this shortage into catastrophic proportions. There is a purely quantitative aspect to this problem; literally tens of thousands of personnel are needed, but only a few hundred are being trained each year. Current efforts to increase the numbers being trained are confined mainly to term institutes and workshops.

But there is also a qualitative problem. The report of the "Roles for Researchers" project[6] currently being concluded at Indiana University shows that the kinds of persons needed are not likely to be developed by existing training programs that have either the flavor of educational psychology or of the traditional tests and measurements. There is, moreover, no agreement about the nature of the emergent evaluator role. So for example, the director of a particular Research and Development Center has said, "We are having trouble finding people who come to us with sufficient sophistication so that they can help with technical problems. We need an evaluator

---

[6]David L. Clark and John E. Hopkins, "Roles for Researchers," CRP Project No. X-022, Indiana University, in progress.

interested in measuring change, who is statistically competent and has all the characteristics of a stereotype methodologist in evaluation but who has a willingness to look at new kinds of problems." The model of the evaluator being developed by the Pittsburgh Public Schools has a definite linkage to the entire change process mechanism in use in that system, so that the evaluator is in fact a kind of change agent. In other instances the evaluator role is defined in terms of competence in a discipline first, and technical skills second. There is thus no consensus, and there are certainly few places where persons are being prepared systematically in these new orientations.

Thus we are faced both with the lack of persons who can function in evaluator roles and with the lack of concepts and materials that are necessary to train recruits into the profession.

Where Next?

I have with malice aforethought painted a rather dismal picture of the state of the evaluative art. Surely the seven lacks that I have described (which are only the most major among literally dozens that might be identified) pose a formidable challenge to the professional community. Even the best evaluators can function only with extant theory, concepts, designs, tools, mechanisms, and training. The practitioner can hardly be blamed if, when placing his faith on those extant procedures and consultant help, he produces evaluative results of little use to anyone. Nor can we fault him too much if he becomes disenchanted with the substitutes we offer because they are not operational.

The primary task in evaluation today is the provision of sensible alternatives to the evaluator. The evaluation of educational innovations awaits the modernization of the theory and practice of the evaluative art.

Is there any hope that this modernization will occur soon? I believe that there is a great deal of reason to be hopeful. Some of the reasons will become apparent, I am sure, after you hear the propositions to be put forth by my colleagues on the panel. We can allude to others briefly here; for example:

On the matter of definition, a number of fruitful efforts have already been made. Cronbach, Stufflebeam, Scriven, Stake, Pfeiffer, Suchman, Quade, and others have assayed new formulations that are somewhat convergent. The national Phi Delta Kappa panel convened for the purpose of writing a monograph on evaluation have pulled these definitions together into a highly useful version that links evaluation and decision-making.

On the subject of decision-making theory, the work of Braybrooke and Lindblom already referred to, together with that of Simon, Hock, and Ott have added useful dimensions to our thinking.

In relation to values and criteria, Quade, Kaplan, Bloom, Krathwohl, and Clark and Guba have made significant contributions.

In relation to data processing (particularly in the form of data banks) and the levels problem, much can be gleaned from the experience of Project Talent, the Measurement Research Center at the University of Iowa, National Assessment, and Project EPIC. Computer capabilities unknown a few years ago also adds a dimension.

In the area of methodology we can look to developments such as quasi-experimental design, convergence technique, Delphi technique, item sampling, Bayesian statistics, PERT, operations research techniques, systems analysis, and the like for some new insights.

Thus the picture is by no means all drawn in shades of black or gray. The profession does show many signs of awareness to the problems that I have described. What is important now is that these first efforts be vigorously pursued and made operational as quickly as possible.

Ladies and gentlemen, the challenge is before us. How will you respond?

Daniel L. Stufflebeam
February 1969

AN EMERGENT THEORY OF EVALUATION

Dr. Guba has attempted to validate the need for a new theory of educational evaluation. In my ten minutes, I will briefly describe some of the results, to date, of a three year effort to develop such a new theory.

Largely this effort has been conducted by the Phi Delta Kappa National Study Commission on Evaluation.*

To develop a new evaluation theory it is necessary to address many difficult questions. Among these are the following:

What premises are fundamental to the theory?

How should evaluation be defined?

What steps are involved in carrying through an evaluation?

What kinds of questions should evaluation studies answer?

What kinds of designs are required to answer these questions? And,

What criteria are appropriate for judging evaluation studies?

Subsequent papers in this symposium will deal with the issues of evaluative questions and designs. This paper focuses on the other four issues, i.e., premises for a new theory, a new definition of evaluation, the steps in the evaluation process, and criteria for judging evaluation studies. Without further introduction let us consider each of these topics.

Premises

Thus far six premises have been identified to undergird the emergent theory. They are as follows:

---

*Members are Walter J. Foley, William J. Gephart, Egon G. Guba, Robert L. Hammond, Howard O. Merriman, Malcolm M. Provus, and Daniel L. Stufflebeam.

1. The purpose of evaluation is to judge decision alternatives; to evaluate, it is therefore necessary to know the alternatives to be judged and the criteria for judging them.

2. To apply criteria to decision alternatives it is necessary to have relevant information; thus, the theory of evaluation must incorporate information theory.

3. Different settings require different evaluation strategies; therefore, the new theory should distinguish between different educational settings and evaluation strategies.

4. Different decision questions require different evaluation designs; therefore, an efficient evaluation theory should define different types of decision questions and corresponding types of evaluation designs.

5. While the substance of different evaluation designs varies, a single set of generalizable steps can be followed in the design of any sound evaluation.

6. Since evaluation studies should answer decision-makers' questions, evaluation designs should satisfy criteria of practical utility as well as criteria of scientific adequacy.

Evaluation Defined

Given these six premises it is proposed that evaluation be defined as follows:

EVALUATION IS THE PROCESS OF DEFINING, OBTAINING, AND USING INFORMATION TO JUDGE DECISION ALTERNATIVES.

There are three things to note about this definition.

First, it portrays evaluation as a process. Process is defined here as a continuing, cyclical activity subsuming many methods and involving a number

of sequential steps. This dynamic, complex conception of evaluation as a recurrent process is in sharp contrast to the relatively static, terminal, single-phase conception of evaluation that is current.

Second, this new definition divides the evaluation process into three parts.

The first part involves defining the information to be collected. The second part pertains to obtaining the information. And the third part pertains to using the obtained information.

The final thing to note about this new definition is that the purpose of evaluation is to provide information for decision making. To evaluate, the decisions to be served should be known in advance. Thus, the evaluator must be a student of the decision-making process.

To reiterate:

EVALUATION IS THE PROCESS OF DEFINING, OBTAINING, AND USING INFORMATION TO JUDGE DECISION ALTERNATIVES.

Steps in the Process of Evaluation

Given this definition, let us consider the evaluation process. It has already been noted in our definition that this process has three steps: defining, obtaining, and using information. Each of these steps will be considered separately.

1. Defining Information Requirements.

The first step in the evaluation process is that of defining. Its purpose is to specify the decision situations to be served, the system within which the evaluation is to occur, and the policies which will

govern the evaluation. The essence of the definition step is to explicate the decision alternatives of interest, and the criteria for judging them. In doing this it is necessary to determine who the decision-makers are, what decision questions should be answered, when the decisions have to be made, what alternatives will be considered, what criterion variables are important and what standards will be applied with each criterion variable. Clearly, definition is the fundamental step in the evaluation process. If it is done poorly no amount of rigor in the data collection and analysis operations can help.

2. Obtaining Information

The second major step is to obtain information. This step must be keyed closely to the criteria and to the alternatives which were identified in the defining step. So for example, if cost is a criterion, one must be sure to collect cost information for each of the alternatives under consideration. Essentially, the obtaining step is the information specialty step. This step includes all of the operations in collecting, organizing and analyzing information. To obtain information one must therefore pay attention to sampling, instrumentation, data collection, information storage and retrieval and statistical analysis.

3. Utilizing Information.

The third step in the evaluation process is the utilization of information. This step provides the decision-maker with timely access to the information he needs. Also it should provide the information in a manner and a form which will facilitate a decision-maker's uses of the information. In accordance with the policy for evaluation, audiences

for evaluation reports should be defined. Appropriate information should be provided to each audience. And the audiences should be assisted to use the information to make decisions.

Criteria for Judging Evaluation Studies

This concludes the description of the evaluation process. Next, let us consider briefly how the evaluator can evaluate his own activity. The information the evaluation produces is the key. What criteria are appropriate to this information?

This question can be answered in two parts. If evaluation produces information, then that information must meet criteria that are ordinarily required of any good information, i.e., scientific criteria. But because it is evaluative information it must also meet criteria of practical utility. Let us briefly consider these kinds of criteria.

The scientific criteria include internal validity, external validity, reliability, and objectivity. Since these criteria are well defined in the literature of educational research I shall not describe them further.

In addition to the scientific criteria, seven utility criteria should be met by evaluative information. These are: relevance, significance, scope, credibility, timeliness, pervasiveness, and efficiency. Let us briefly consider each of these.

To be relevant the information must relate to the decisions to be made.

To be significant the information must be weighted for its meaning in relation to the decision. Not all relevant information is equally weighty. The culling and highlighting required is a professional task that justifies the inclusion of a reporting expert on the evaluation team.

To have adequate scope the information must relate to all aspects involved in the decision. If there are six alternatives to be considered, information that pertains to only four lacks scope. The same may be said if some of the specified criterion variables have not been considered.

To be credible information must be trusted by the decision-maker and those he must serve.

To be timely the information must come in time to be useful to the decision-maker. The evaluator must guard against the scientific value that argues against publishing findings until every last element is in. Late information is worthless information. It is better in the evaluative situation to have reasonably good information on time than perfect information too late.

To be pervasive the information must reach all of the decision-makers who need it.

To be efficient costs for evaluation must not mushroom out of all proportions to its value. The imprudent evaluator may produce a mountain of information whose collection imposes an intolerable resource drain. Proper application of the criteria of relevance, significance, and scope should remedy the grossest inefficiencies. But even when the information proposed to be collected meets all of these criteria, there are probably still alternative ways for collecting it that differ in terms of the resources that are required. The criterion of efficiency should guide the evaluator to the appropriate alternative.

An evaluator who can say, after careful examination, that his evaluation design will produce information that conforms to all of the scientific and utility criteria can be assured that he is doing his job well.

Finale

This concludes my presentation. Due to time limitations my remarks have been cryptic. I hope that I have not confused you too much. I do hope, however, that you have been stimulated to think about the difficulties inherent in projecting a theory and a methodology of evaluation which are at once scientifically respectable and useful to practitioners.

KNOWLEDGE ABOUT DECISION PROCESSES AND INFORMATION

Robert S. Randall

Director
Division of Program Research and Evaluation
Southwest Educational Development Laboratory
Austin, Texas

A paper read at the symposium on
"The World of Evaluation Needs Reshaping,"
The American Educational Research Association Convention
Los Angeles, February 1969

INTRODUCTION

There is a timeworn and oft-recurring spectacle of the frantic but finally productive researcher-evaluator, who rushed into the executive offices with his data analysis finally complete, his report prepared and in hand, only to find that the executives, several months previously, had made the important decisions that locked up the monies and committed the organization for the ensuing months ahead. This illustrates the tragic failure of evaluators and evaluation systems to focus attention on the nature of decisions and the time when they are to be made. As we gain new knowledge about evaluation and its effects on programs and funding, it becomes patently clear that attention must be focused on decisions, their nature, when they are made, and the information needed on which to base them. Until evaluators come to grips with this central issue, we will likely continue to produce reports that have little effect except on other evaluators and researchers (and of course on students who write theses). Let us examine some of the problems of evaluation as they relate to decisions.

ANALYSIS OF THE EVALUATION PROCESS

For purposes of this paper, evaluation is defined in general to be the process of choosing among alternatives while utilizing the best information that is available. This definition puts the emphasis on valuing, but valuing based on sound, relevant information. A more specific definition of evaluation is the process of maximizing the effectiveness of decisions through the timely reporting of relevant information in a useful form to appropriate levels of decision-making. This means that both key decisions and the time they will be made are identified as a requisite to identifying, collecting, analyzing, interpreting, and reporting the relevant information. It must be

clear that the most reliable and valid information is almost useless if it arrives too late to be considered.

This notion of evaluation gives rise to several important questions. Who influences as well as who makes decisions? What is the nature of the decision to be made? What are the constraints and criteria that affect the decisions? What is the nature of other information on which the decision might be based? When will the decision be made? Can it be postponed? An adequate evaluation system must seriously consider effective ways of responding to each of these questions.

The CIPP Evaluation Model, developed by Stufflebeam and Guba, attempts to take into account such factors. Four classes of decisions are postulated in the model. These are called planning decisions, structuring decisions, implementing decisions, and recycling decisions. The relationship among these decisions, information on which they are based, and the sources of information are illustrated in Figure 1. Let us examine each class of decision, looking at the state of the art in terms of how much knowledge is available about the decision process and the information requirements.

Planning Decisions

Planning decisions involve setting priorities in terms of problems to be attacked, and selection of a strategy or strategies through which the problems might be attacked. Such decisions are usually made at or near the policy level in an organization. Educators have often made such decisions "off the top of their heads," and it has been unusual when anything other than sporatic or haphazard analysis has had effective influence on such policy decisions. Economists and philosophers have long proposed idealistic, analytical models based on a rational-deductive system to be used in making policy decisions, but these models have proved to be of limited use. How-

Figure 1

Decision - Information Matrix

Decisions	Kinds of Information	Source of Information
Planning	**Context**	
Problem Selection	Organization Constraints, Nature of Conditions, Setting	Policymakers, Research Surveys, Experts
Strategy Selection	Resource Constraints, Criteria, Alternative Strategies, Methods, Approaches	Funding Sources, Reported Research, Experts
Structuring	**Design**	
Component Objectives	Tested goals, Theory, Models	Research reports, Experts
Component Activities	Tested procedures, Educated guesses, Intuitive hunches	Research reports, Experts
Restructuring	**Process**	
Component Objectives	Effectiveness evidence	Subjects, Participants
Component Activities	Practicality of use, Tested Procedures	Participants, Research
Recycling	**Product**	
Multiple Components	Comparative Effectiveness evidence	Observation and Testing of subjects, controls, and/or methods

ever, recently some breakthroughs have been made in studying how such policy decisions are made. Most notable is the work of Braybrooke and Lindblom (2), a philosopher and economist, who combined to describe what they termed "disjointed-incrementalism." Their system relieves the decision-maker of the impossible burden of considering all possible alternatives and their consequences in making policy decisions and proposes rather that policy decisions be made on a basis of taking incremental departures from the situation as it exists. After effects of the increment are noted, other incremental steps can be planned. This approach is further elaborated by Lindblom in a paper called, "The Science of Muddling Through," (4) which I commend to your attention.

However, even in incremental policy decision-making, certain information needs are apparent. It is important to understand the nature of the situation that exists. Survey research methods are useful in this effort, but information retrieval becomes a problem. Although some efforts are now being made to establish data banks and new information processing and retrieval systems, they have not been developed to the extent of being entirely useful to planners. However, the state of the art in studying policy decision-making and information systems that will yield data on which to base such decisions is far ahead of the studies of other kinds of decision-making and information needs.

Structuring Decisions

A second class, structuring decisions, entails choosing among alternatives in producing designs. Extensive study and theory development in decision-making by Barnard (1), Simon (5), and Griffiths (3) has focused attention primarily on decisions that are made in the course of operating or

maintaining an organization. While models they have proposed are of use in studying administrative decision-making, they are little help to those who consider the nature of decisions made in producing or choosing among alternative designs. In fact, a search of the literature makes it appear that designers and those who choose among designs are presumed to be gifted with some superhuman guidance that enables them to determine intuitively and on examination the design that is most adequate. We are in dire need of new knowledge and new study about how such decisions are made or how they might be made. Accordingly, knowledge about information needs is scarce. Some help is available from educational and psychological measurement studies for facilitating information systems, but the information needs have not been sufficiently well studied to determine the requirements adequately.

Implementing Decisions

A third class of decisions is implementing or restructuring decisions. As a new design is put to the test, it is assumed that it will have some defects and will need some restructuring. Therefore, the assumption is that some things can be learned during the test which will enable the designer to refine and modify the plans and procedures. The questions are: What is the nature of the decisions he makes, and what information will influence him in making these decisions. Here again the theories of administrative decision-making are of some use in organizing and facilitating communication but they are of little help in analyzing or discovering the nature of restructuring decisions based on information processing. In addition, information theories are lacking. Classical research designs impose unrealistic and undesirable constraints upon the information process needed while a program is being tested and redesigned. Hence, studies and new theories of decision-making and information needs, relative to implementing and restructuring designs, are sorely needed.

Recycling Decisions

The fourth class of decisions postulated are called recycling decisions. Such decisions determine whether to continue, terminate, modify or refocus a project. The decisions depend on information about attainment of stated objectives and comparisons of effects with those of other methods. Much of the work in psychological and educational measurement and design is most appropriately applied to product evaluation. However, getting the information in a timely manner has always plagued evaluators and project managers. Hence, more effort is needed by researchers and evelopers on obtaining and displaying information in the time and form that decision-makers require. Now let us turn attention to some operational difficulties that affect decision and information process.

OPERATIONAL DIFFICULTIES

Having discussed the nature of decision and information needs we have some operational problems are noted. A great deal of difficulty is encountered in identifying decisions and decision-makers and the information that is relevant to their decisions.

Identifying Decisions

Decisions that are faced are not always easily recognized. Often decision-makers themselves are not fully aware of the decisions they may face. In introducing new information, the evaluation system may focus attention on decisions that were not previously considered. Hence, the system must provide persons who are in contact with key decision-makers and are continually alert to decisions that will be faced.

Another problem in identifying decisions and their nature is that decision criteria may change as time passes. New developments occur; new information is obtained; conditions change as time goes by. Any one of

these can cause new criteria to appear or old ones to be of no effect. Hence, the system must provide for a continual reassessment of criteria that may affect decisions.

The passing of time may also cause constraints to change. Since there is always some lag between the time when decisions are identified and the time when information is collected, processed, and reported, the system must continually be alert for changes in constraints that might change the basis on which decisions will be made.

Identifying Decision-Makers

Another problem is the identification of persons involved in the decision process. These include not only those who have final authority in making decisions but others involved in the decision process who may influence the final decision-maker. Typically, the decision process in an organization involves a complex network of persons who have varying degrees of influence on the one who may have constituted authority to make any given decision. Hence, it may be useless to get information to the recognized, final decision-maker, in that he either may have little time for considering the information or may rely heavily on the judgment and recommendation of other people. Therefore, the evaluation system must identify the key persons involved in any strategic decision and make arrangement for getting necessary information to these people.

Timing of Decisions

The best information is of utterly no use if it does not arrive in time to base a decision on it. Therefore, the key for the operation of an evaluation system is to get the best information possible in the time that is allowed. Of course, it is possible to postpone the time of the decision,

but often such a delay is not possible. Hence, the system must respond to the time when critical decisions will be made and yield the information needed in time for it to be considered.

Identifying Relevant Information

It is not enough for evaluators to decide what information would be best on which to base the decision. Cues must be taken from the decision-makers as to what information is relevant to their decision tasks. It is useless to force sophisticated information upon a decision-maker who fails to see its relevance, since he will ultimately disregard it in favor of more understandable, if less relevant information. The system can be designed to educate decision-makers to the usefulness of certain kinds of information, but the final criterion must be that the decision-maker considers the information relevant. Otherwise, the best information will have little, if any, effect on the decision.

Reporting in a Useful Form

Another problem related to the relevance of information is to get the information to the appropriate decision-makers in a form that is most useful to them. This entails not only varying the degree of sophistication but also the degree of specificity of reports. The criteria must include the length of the time the decision-maker will likely have to consider the information as well as his competence in understanding the terminology and techniques used to present the information. Thus, the same information may be presented in several different forms to different decision-makers.

It is obvious from the preceding discussion that communication and interaction with key decision-makers is a cornerstone on which effective evaluation rests. We tend to make many unwarranted assumptions about the effectiveness

of our communications. One of the hazards of written communication is that the writer has little control over who will read his paper, what psychological set they will have as they read it, or how they will interpret it. Furthermore, he has no chance to interact or clarify his meaning or intent with many of the readers. Therefore, the more visual and oral cues and face-to-face interaction that can accompany his written communications, the more chance he has of being understood. Such research as we have on communication suggests that we are more likely to fail to be understood than to communicate effectively if we depend on any single sensory perception.

SUMMARY

This analysis has tried to show that while some efforts have been made to study decision processes, and methods of obtaining, storing and retrieving information, a void still exists. Knowledge of the decisions and information processing needs for <u>context</u> and <u>product</u> evaluation is barely adequate, but huge gaps exist in the knowledge of the nature of decision and process information needs for effective <u>input</u> and <u>process</u> evaluation. In addition, operational difficulties in identifying decisions that are faced, those who effect them, their timing, and reporting relevant information to the decision-makers in a timely and useful manner are factors that threaten to sabotage the efforts of the best intentioned evaluator. Thus, our needs are great. But recognition of need is the first step toward solution of the problem. It is hoped that this discussion may induce some of you to discover, develop, or inspire the development of some of the required knowledge.

REFERENCES

1. Barnard, Chester I., The Functions of the Executive, Cambridge, Mass., Howard University Press, 1938.

2. Braybrooke, David, and Charles E. Lindblom, A Strategy of Decisions, New York, The Free Press, 1963.

3. Griffiths, Daniel E., Administrative Theory, New York, Appleton, Century, Crofts, Inc., 1959.

4. Lindblom, Charles E., "The Science of Muddling Through," Readings in Managerial Psychology, H. S. Leavitt and L. R. Ponay (eds.), Chicago, University of Chicago Press, 1964.

5. March, James G., and Herbert A. Simon, Organization, New York, Wiley, 1958.

Symposium "The World of Evaluation Needs Re-shaping"*

Evaluation Designs and Instruments
Jack C. Merwin
University of Minnesota

I was happy to accept the chairman's invitation to participate in this symposium because I felt the title reflected many of my personal biases. Within the framework of our frustrations with available designs and instruments which do not meet many of our varied needs for evaluation, the term re-shaping implies to me, 1) consideration of where and how currently available theories, designs and instruments are proving useful, 2) identification of needs that cannot be met with currently available constructs and tools, and 3) an attempt to identify guidelines for efforts to meet unfulfilled needs.

In my brief comments this morning, I will attempt to put the dimensions of our current needs in a historical perspective. The most promising aspect of current frustration is the long overdue recognition that we can no longer live with the totally unrealistic idea that a small number of designs and a very limited variety of evaluative instruments can serve all of our needs for evaluation in education.

I view the following as encouraging signs of movement and trends toward the needed reshaping of the world of evaluation as it relates to evaluating individuals:

---

*Annual meeting of the American Educational Research Association, February 1969

1. Emphasis on measuring change, rather than status, many problems of which are brought out in a report of the Wisconsin Symposium, Problems in Measuring Change, edited by Chester Harris.
2. Explorations of the use of sequential procedures for gathering information, as opposed to across the board administration of instruments.
3. Experimentation with placement tests, "imbedded" items and proficiency tests as part of the learning process, such as that of the Oakleaf Project of Glaser and his associates.

On the latter of these points, it is interesting to note something similar from the past. Monroe's book of 1918, Measuring the Results of Teaching, carried a focus on mastery of skills related to very specific objectives.

Our evaluation efforts in recent decades have focused on evaluation of the individual and indeed there is further development and reshaping needed in this area. But there have been other needs for evaluation which have gone largely unheeded for some time. In his paper on Course Improvement Through Evaluation, Lee Cronbach describes the situation in this way:

> Many types of decisions are to be made, and many varieties of information are useful. It becomes immediately apparent that evaluation is a diversified activity and that no one set of principles will suffice for all situations. But measurement specialists have so concentrated upon one process--the preparation of pencil-and-paper achievement tests for assigning scores to individual pupils--that the principles pertinent to that process have somehow become enshrined as the principles of evaluation.

Much recent concern has not been with evaluation of individuals but with evaluation of programs; instruction, curriculum, methodology and so forth. Looking to the past first, we note that at the turn of the century there was a similar concern. Rice's classic study of the 1890's was aimed at a comparison of outcomes of different approaches to teaching the same subject. The 1916 NSSE Yearbook was entitled Standards and Tests for Measurement of the Efficiency of Schools and School Systems. That same year, Arnold produced a book entitled Measurement of Teaching Efficiency. In 1918, Monroe authored a book entitled Measuring the Results of Teaching, and the NSSE Yearbook for that year was The Measurement of Educational Products. It was with the background of design and instrumentation set forth in such books that the great expansion of achievement testing took place in the 1920's.

I believe Cronbach hit upon the basic reason for many of our frustrations today as we look to currently available designs and instruments for program evaluation. He wrote,

> At that time [1920], the content of any course was taken pretty much as established and beyond criticism save for small shifts of topical emphasis. At the administrator's discretion, standard tests covering the curriculum were given to assess the efficiency of the teacher or the school system. Such administrative testing fell into disfavor when used injudiciously and heavy handily in the 1920's and 1930's. Administrators and accrediting agencies fell back upon descriptive features of the school program in judging adequacy. Instead of collecting direct evidence of educational impact, they judged schools in terms of size of budget, student-staff ratio, square feet of laboratory space, and the number of advanced credits accumulated by the teacher.

In this article from the Teachers College Record in 1963, Cronbach's next sentence is "This tide, it appears, is about to turn." Today we are

looking at the needs for evaluation designs and instruments from a somewhat different view than our predecessors of the 1920 era. We are concerned not only with effectiveness of teaching, but also the effectiveness of "innovations" in all aspects of education.

Since the 1930's testing has been almost exclusively designed for judgments about individuals. Summary figures across scores for individuals have provided some information regarding program effectiveness. We have been all too long, however, in coming to the realization that this approach often is not only inefficient, but simply does not provide some of the information needed. Thus, whether we attribute it to requirements for evaluation written into federal legislation, new approaches to teaching, or numerous curriculum development projects, the pressure has mounted to produce what I consider to be a healthy concern about the need for reshaping evaluation methodology and instruments to implement that methodology.

Irritating as it is to face broadened evaluation needs and find that available tools will simply not do the job, several types of activity already started indicate movement in promising directions.

One such activity that I would cite is the proposed use of a decision-making framework as a basis for thinking about evaluation. Stufflebeam has been working specifically on educational decision making as a framework, and Cronbach and Glaser earlier had set forth a general background. Stake's paper, "The Countenance of Educational Evaluation" provided a refreshing new view. The attention being given to mastery testing by Glaser et al at Pittsburgh and Bloom in Chicago, along with

the work on "Universe-defined" tests by Osborne and by Hively have been interesting new developments. Cronbach's proposal for an unmatched design for collecting information from groups should be included in this list, as should the efforts toward unique designs and instrumentation that has been under development by the Committee on Assessing the Programs of Education. And, I should not end this listing without mentioning the AERA Committee on Curriculum Evaluation and the monograph series started by that Committee.

I also want to mention some concepts of relatively recent vintage that have not bean in the focus of design and instrument development, but which may well help us in reshaping of the world of evaluation around design and instrumentation. One is the distinction between formative and summative evaluation set forth by Scriven. A second is the concept of fidality versus bandwith of information suggested by Cronbach and Glaser. A third is the general idea of group evaluation as opposed to individual evaluation. And, finally, I would propose that all of such concepts might most readily move us toward a positive reshaping of evaluation if our needs for evaluation can be examined within the framework of educational decision making.

THE WORLD OF EVALUATION NEEDS RESHAPING

by

Dr. Michael C. Giammatteo

Paper Presented at
AERA Symposium: "The World
of Evaluation Needs Reshaping"
Los Angeles, California, February 1969

Within the last decade or so the scope of the educational researcher has greatly enlarged; it has also undergone some very decisive changes. The other members of our panel* have described these concerns which I must confess are exciting. These expanding concepts will demand a receptive audience in terms of both researchers and users of research services. As the educational researcher's role is expanded in scope and sophistication of technique so must his training. Not only what people do, but what they intend to do, and what they expect to happen are now objects of systematic analysis. When the location, examination and nature of data changes, history changes its character. When rapid feedback of data reduced to information occurs we also change the character of history. Educational researchers are too vital to be trained only as technicians.

I propose that training of this body of professionals--researchers commence with preservice education at the junior level--continue through the on-the-job level, and continue at the inservice level above and beyond the doctorate.

The major clusters around which the theoretical thinking and research operations cluster are as follows: (illustrative not exhaustive)

Cluster I - Cultural Blocks

Major foci:

Study anthropological approaches

Focus on primary message units

*Egon G. Guba, Director, National Institute for the Study of Educational Change, Indiana University; Daniel L. Stufflebeam, Director, Evaluation Center, The Ohio State University; Robert S. Randall, Associate Director, Division of Research and Evaluation, Southwest Educational Development Laboratory; Jack C. Merwin, Director of Psychological Foundations, College of Education, University of Minnesota.

Major foci: (Cont.)

Focus on interaction between technical, formal and informal systems*

Language analysis

Ecology

Proponents/Referents: Ruth Benedict, Edward Hall, Muriel Hammer, Claude Levi-Strauss, Owen D. Lattimore

Cluster II - Agency Entry and Interfacing Roles

Major foci:

Decision structures

Collaboration

Adaption

Utilization of knowledge

Linker roles

Profile Development**

Content analysis for mass communications

Target group analysis

Proponents/Referents: Ron Lippitt, Henry Bridell, Everett Rogers, David Clark, Knowledge Utilization Center, Northwest Regional Educational Laboratory, Egon Guba, Braybrooke

Cluster III - Problem Solving Roles

Major foci:

Communication skills

Interview skills

Data reduction

Force field analysis

Creative problem solving

*See Map of Culture (Table I - Page 7)

**See A Model of the Adoption of an Innovation by an Agency Within a Contextual System (Table II - Page 10)

Proponents/Referents: H. Thelen, R. D. Laing, H. Phillipson, A. R. Lee, American Management Association, K. Lewin

Cluster IV - Issue Analysis Roles

Major foci:

Political analysis

Trends analysis

Polling

Issue analysis

Gaming and simulation

Indices development

Proponents/Referents: Clark Abt, Don Oliver, Shaver, E. Fenton, Donald M. MacKay, Anatol Rapoport, David Easton, and Mervyn L. Cadwallader

Cluster V - Management/Systems Techniques

Major foci:

Program evaluation and review techniques

Critical path modes

Time-cost-performance factors

Cybernetics

Time lines

Network

Topology

Graph theory

Proponents/Referents: D. Cook, Ken Boulding, L. Von Bertalaniffy, N. Wiener, Military

Cluster VI - Information Science (retrieval techniques)

Major foci:

Information theory

Time lags

Major foci: (Cont.)

Timeliness

Laws of requisite variety

Content validity

Denial systems

Data processing and retrieval

Abstracting and indexing

ERIC

Proponents/Referents: G. A. Miller, W. Ross Ashby, Anatol Rapoport

Cluster VII - Measurement

Major foci:

Instrumentation

Reliability

Validity

Clinical

Experimental

Observational

Proponents/Referents: N. Gage, Charters, Best, Ebel, Stack

Cluster VIII - Models/Tools/Techniques

Major foci:

Standard statistical treatments

Philosophy of science

Design concerns

Models and paradigm from anthropology

Sociology

Psychology

Economics

Industry

Proponents/Referents: J. Stanley, Campbell, W. Borg, Lindquist, Edwards

The illustrative entries will give you some flavor of the types of skills that might be in the experiences of the researcher.

The following discussion cites a walk through of one major concern confronting today's researcher. Namely that of interfacing with multiple agencies. Here we have some of the problems a researcher working with a Title III project group faces.

The basic problems to be faced in replicable programs of planned change in educational practices are firmly rooted in the cultural setting. Besides the technological revolution, informational and the human rights revolutions have occurred at a more rapid rate than have changes in the educational and behavioral sciences. The cultural setting poses many problems which have implication for designers of programs of planned change as well as the researcher. The two major problem areas center on these two foci:

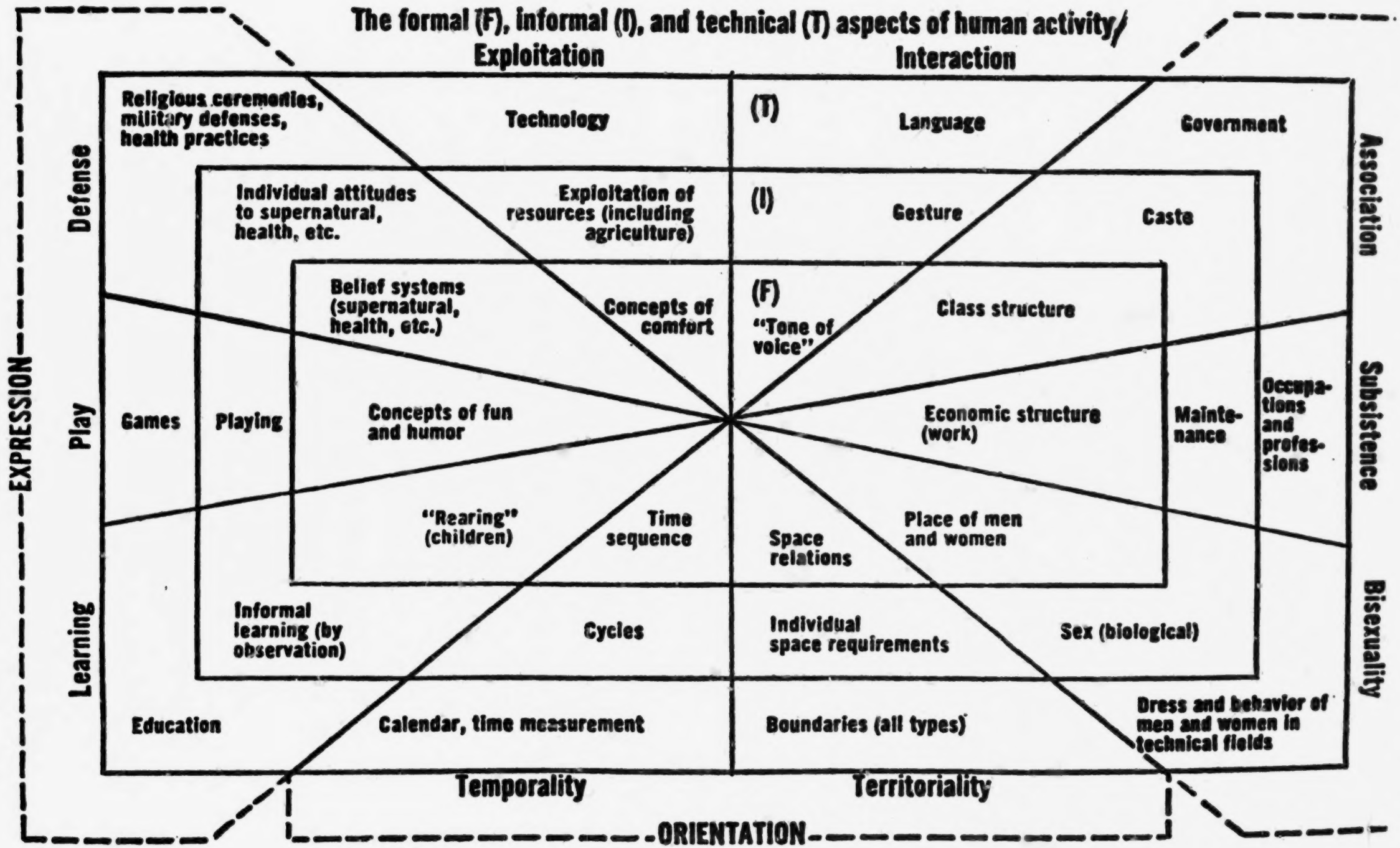
1. What knowledge and skills will be needed for initial effective entry into the adult world?
2. Which knowledge and which skills are the responsibilities of which educational group? (i.e., business, industry, defense, public schools, the home, church, etc.)

In earlier presentations by colleagues it became clear they were focused on the researchers' concerns. However, the contextual systems where research occurs is user oriented, not research oriented.

The first problem relates to educational practices being firmly rooted in the cultural setting. Table 1, a Map of Culture, is based on Ruth Benedict's work in the early 30's and more currently by Dr. Edward Hall's works in the last decade. The following discussion relates to Table 1. The outer ring where the (T) is located describes by title the kinds of systems in the culture that are replicable many miles away. These are given a technical (T) status. For example, language. If we talk about a language phenomenon, e.g.,

Table I

The formal (F), informal (I), and technical (T) aspects of human activity



the use of a question mark, that discussion related to a group of other people familiar with that technical system (language) will be able to employ this information. The information is communicable over great distances with maximum assurance it will be understood by the receiver. The same is true for mathematical concepts. If it is technical and we can transmit the message over great distances without the specialist there, we have satisfied the definition for something technical (T). Government is technical in that the laws can be transmitted a great number of miles and understood. Notice we are saying nothing about the interpretation of the message, only that the message can be understood. In mathematics if you send the mathematical formula x = y many miles away, it will be understood, e.g., calendar and time measurements.

The ring where (I) appears includes informal kinds of cultural barriers. The best trained technician in the mathematical field might be unable to break through and teach because of his informal types of behavior. As a result, whether he is dealing with an individual or an agency, he will not be able to affect change in that situation. In other words, a highly competent mathematician couldn't communicate his technical level skills if his personality did not blend itself well to that informal type of behavior acceptable to the agency. The tone of voice and the concepts of class structure and the other things you see in the inner circle are of prime importance. Let's take our mathematician friend, and assume he understands all the technical items in the mathematical language. Even if he has the proper informal behavior in the atmosphere around him when he tries to explain his technical understanding, he might encounter resistance in the receiving person or agency. If his concept of childrearing is so uniquely different from his client's he may be suggesting ideas so foreign as to be rendered totally unacceptable. That is to say, his form of childrearing

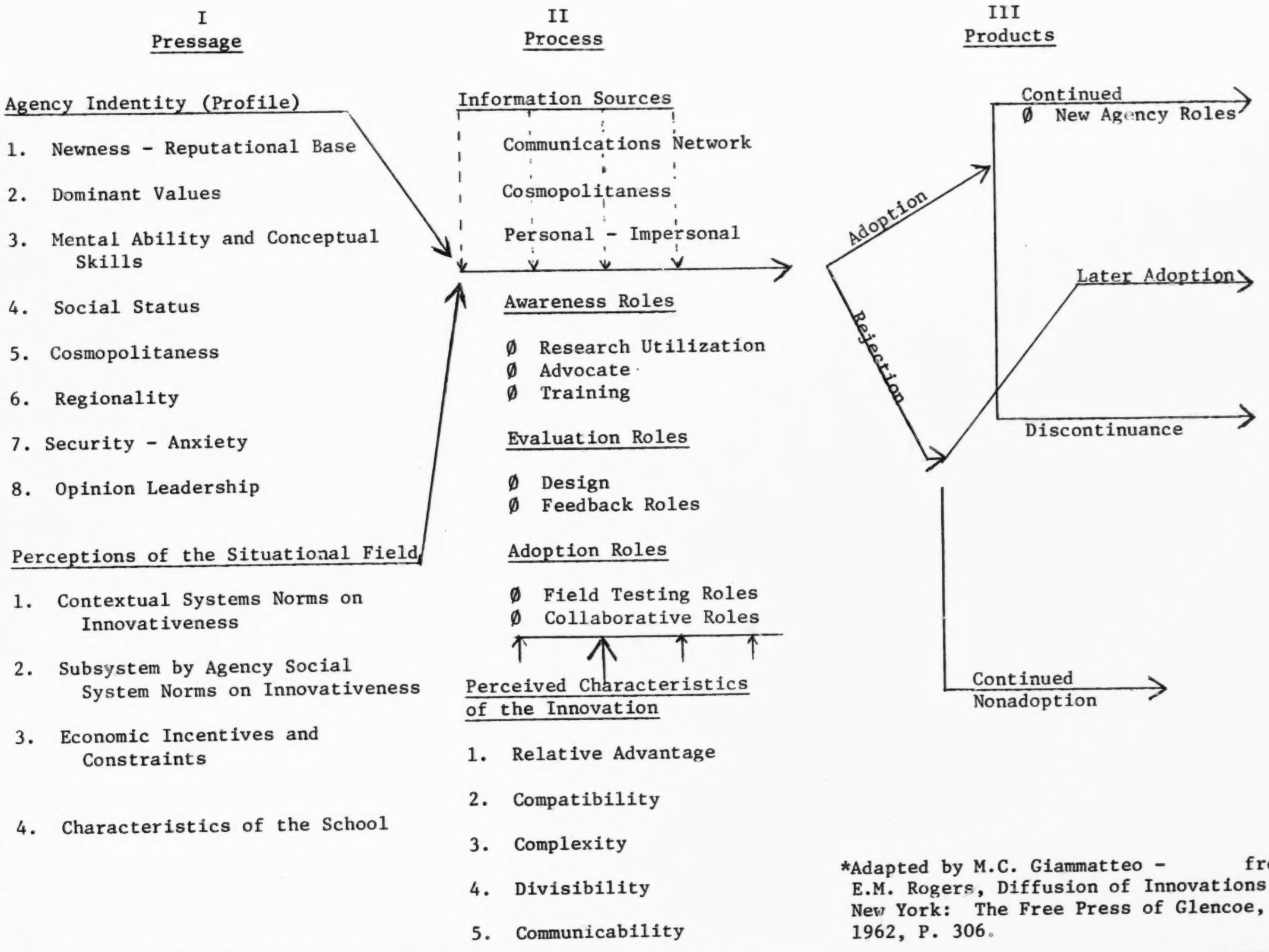
was that the child should be seen and not heard and he happens into a community where the mental health view was held in high esteem (that is, the child seeks and becomes an inquirer and explorer). Our mathematician may find that his competence and technical training do not help him break the barriers because he is violating informal systems. For example, if we trained a white person in the best formalized techniques in education dealing with the phenomenon of reading and we trained him in all the informal behaviors that exist in a locale we may still face failure in our research/ change effort. If this person's formal concepts are violently different to that of the formal system of the place where this man will practice the art of teaching reading, he will have encountered a cultural barrier. The obvious list of variables jumping through your minds demands we recognize the cultural setting in cross agency research. We are talking about things which are much more deeply seated. Indeed, if you progress from the outer ring of technological systems to the inner ring of formalized systems, you will soon begin to develop your own concepts of what constraints you encounter in cross agency work. It is a hope that Table 1 and the above discussion will aid you to understand why the very human level resistances occur.

Point one of this portion of the paper is to recognize the cultural setting must be understood by the researcher. Please do not look at its simplicity. Take an intense look at some of the minimal parameters as suggested by the Map of Culture.

The second table offers a construct developed by Rogers in his excellent book, "Diffusion of Innovations." Column I - Pressage Concerns, has eight major items that permit you to identify an agency profile. It is important for you as you enter the arena of cross agency work to understand the reputational base the agency is trying to encourage. You may find that it wants to develop a base saying that, "it is there to render services." This

Table II

A Model of the Adoption of an Innovation by an Agency Within a Contextual System*



*Adapted by M.C. Giammatteo - from E.M. Rogers, Diffusion of Innovations New York: The Free Press of Glencoe, 1962, P. 306.

will greatly influence your intervention style. If, however, the reputational base it is trying to develop is one of a "learner scholarly group," they may not welcome any intervention or research that deals with applied portions of the research to application continuum. Dominant values of key agency people are important to you as you attempt to develop an agency profile. You should exercise extreme caution in assessing the dominant values in an agency. For example, if x, y and z, the three top dogs in an agency hold dominant values suggesting they support public relations types of activities, and you have been dealing with l, m, n, o and p from that agency and their dominant values are different, you would be advised to identify the dominant values as being those held by x, y and z. The mental abilities or conceptual skills of the people you are working with in your cross agency work are also important factors. Indeed, if the agency staffs are not conceptually oriented you may want to back off of certain types of research approaches. The social status held by the agency is also a determiner of the style of cross agency research you employ. A high prestige agency may not want any experimental approach. The cosmopolitaness of the agency is important. Differing points of view from many types of people around the nation facilitate research interventions in an agency. The more cosmopolitaness, the more apt that internal confrontations will be created. Most often those confrontations are resolved by research techniques. Regionality is crucial. If you deal in a rural, small school setting, it would be ill advised to use certain types of research approaches you might use in another setting. For any of you who have done research in rural isolated areas the concept of regionality is real and live. If you have tried to do sociological studies in the deep south, the fact of regionality becomes of prime importance. The security through anxiety dimension is another area where you must take caution. If the agency is a new one developing a

new reputational base in that region and the individuals in it are highly anxious, they would be ill advised to carry out certain types of research designs. Again, looking at point eight under your efforts to identify the agency, you find that your opinion of the leadership is important to you. The leadership may be different from the dominant values held in an agency. The leader may be so very idealistic that his opinion should be sought out. However, he may be so idealistic so as not to be congruent with the dominant values held by his coworkers and subordinates. Again cross agency work does not permit simplistic answers. Once you have determined that agency profile using the eight entries, you have in essence begun to diagnose your entry. There are instruments that some of you know will permit you to measure or get estimates of each of the eight entries under agency identity.

Now, perceptions of the situation field are crucial because if the contextual system in which all of these things are happening is completely incongruent with the profile you obtain under the eight entry points (under agency identity), then all of the work in the world will not give that agency the kind of feedback it needs to modify its behavior. For example, in some school systems you may do all the agency identification work that is needed. They may ask you to undertake examination of a problem. But if at the very start their perceptions of the situational field differ from the community, you face problems. For example, a black community around a school district may hold ideas that are completely disjointed from the district's. Your actions will probably be blocked, especially if you violate the formal systems. You may even be violating permanent subsystems. For example, if the agency we are identifying above is a school district, one of the subsystems may be considered the school or a school level. If the host agency (the schools) consists of ten schools, one of the schools may be an

all black school. One may be in an area where there are $70-80,000 houses. You had better believe that the subsystem norms on innovativeness will be uniquely different and so the agency titled the school district must be treated as if it were many agencies. Do not fall into the bind of using the school district's profile when you are talking about a particular subsystem.

Point three - Economic Incentives and Constraints. If a situational field implies the district would not financially support an additional research project or implementation strategy noted by the research project, you had better be cognizant of these factors.

Point four, naturally, is the Characteristics of the School - the demographics of the situation. All of these feed into the profile which you must take cognizance of when you are doing research across agencies. Some agencies are trying to provide services under the process area. For example, when we enter into cross agency work, once we have done some of the estimates of the power for and the power against change, based on the dimensions you have seen under column 1, we make an assertion that different kinds of awareness grow and can be facilitated by the intervention agency. Several of the documents we have taken with us here today show how we play the research role, the advocate and training role.* For example, if we work with a school district and we isolate one school that appears to be the constraining school, we find that we must jump to column 2, the process level and facilitate training for it. It is a hope this training may make the school more congruent with the total agency in terms of how it wants to be identified. So there are a number of these process roles you must play prior to completion of any major research function in that setting. We are finding that it is important to sell notions 1, 2, 3, 4 and 5 under Perceived Characteristics of the Innovation. For example, if we can show the school disjointed from the main system the

*Field papers dealing with the clusters are available upon request from the author.

relative advantages for them, we may find they are more willing to participate in the cross agency kind of research. If we showed them that the kinds of things we are doing are compatible with their existing kinds of behaviors, we also may be permitted to carry out the research.

By the way, the research we are talking about here is often confused as the innovation. The complexity of our entry is also vital and if it is too complex, there are certain systems that will not permit you to enter. Some of our interventions in rural settings were perceived of as too complex and thus placed the research effort in jeopardy. Also the divisibility is important. Can we divide elements of the innovation for the research process? For example, in some of our districts we are trying out IPI (Individually Prescribed Instruction). We can divide IPI into subcomponents and just deal with the mathematical package. The effects of the IPI project must be studied in terms of subportions.

Communicability--if part of what we are doing is at a technical level, then we can communicate it a great distance from the test site. The people are sharp enough in the host agencies to know that whatever they do must be refunded, therefore, it must seem to be communicable. You will notice nothing about the replicability problem. The people with whom we are working in our research intervention are not necessarily interested in the same things that interest us in research.

The third column deals with products and these are contributing both to the needs of the research agency and to the host agency. The research agency may use a descriptive analysis as its product. Descriptions of how you enter the system and what you did could be a product, for example, noting how you timed data releases to a host agency so the data could coincide with crucial budget discussion periods. It is the description of the process that is the

product. You may actually develop hardware or software, but as you can see on this chart the concepts that Rogers held are the ones we are holding to. Aside from the several reinterpretations and additions under columns 1 and 2, the basic construct developed by Rogers seems to be one that we should expose to the researchers in training.

Thank you.